Subgroup Discovery in Data Sets with Multi–Dimensional Responses

Lan Umek*, Blaz Zupan†

July 9, 2010

Abstract

Most of the present subgroup discovery approaches aim at finding subsets of attribute-value data with unusual distribution of a single output variable. In general, real-life problems may be described with richer, multi-dimensional descriptions of the outcome. The discovery task in such domains is to find subsets of data instances with similar outcome description that are separable from the rest of the instances in the input space. We have developed a technique that directly addresses this problem and uses a combination of agglomerative clustering to find subgroup candidates in the space of output attributes, and predictive modeling to score and describe these candidates in the input attribute space. Experiments with the proposed method on a set of synthetic and on a real social survey data set demonstrate its ability to discover relevant and interesting subgroups from the data with multi-dimensional responses.

*Faculty of Computer and Information Science, University of Ljubljana, Tržaška 25, SI-1000 Ljubljana. Phone: +386 1 4768933, E-mail: lan.umek@fri.uni-lj.si

†Corresponding author. Faculty of Computer and Information Science, University of Ljubljana, Tržaška 25, SI-1000 Ljubljana, Slovenia. Phone: +386 1 4768402. E-mail: blaz.zupan@fri.uni-lj.si

Keywords: subgroup discovery, multiple responses, hierarchical clustering, subgroup scoring, classification, European social survey

1 Introduction

Subgroup discovery is a data analysis approach that aims at finding descriptions of subgroups of data instances with unusual statistical distribution of the property of interest [21, 37, 22]. Currently prevailing subgroup discovery techniques infer subgroups from class-labeled attribute-value data sets. Prominent approaches in this area, such as EXPLORA [21], SD [7], CN2-SD [24], and APRIORI-SD [20] describe each subgroup through a classification rule. It's condition part, expressed as conjunction of assertions on values of attributes, defines the input attribute subspace where the data instances have an unusual distribution of the class variable. Rules that identify "good" subgroups cover typically a large part of input attribute space where the majority of training data instances are labeled with a single class value. It is desired that predictive accuracy of such rules is high. Similarly to the methods of *predictive induction* [11], subgroup discovery aims to infer rules where such accuracy is attainable. But while the methods of predictive induction construct a comprehensive classifier that preferably performs well on the entire attribute space, methods of subgroup discovery infer rules that cover only a subset of training instances. Instead of maximizing the overall prediction accuracy, they focus on revealing and describing the meaningful data subsets.

Example applications of subgroup discovery include the analysis of clinical data [2, 8, 9], marketing analytics [4], gene expression data analysis [10], analysis of e-learning [31], and analysis of traffic accidents [19]. In each of these applications, the problem was conveniently represented with data in the attribute-value form, and classification rules were used to describe the subgroups. In general,

subgroup discovery should assist us in hypothesizing the relations between a set of input features and observed outcomes. The problems considered by subgroup discovery are often complex, both in terms of representation of input and outcome of the observed system. Representation of the outcome with a single class variable is convenient and fosters the utility of standard data analysis tools. Yet, in many real-life domains, the outcomes are complex and need to be described with a number of features. For instance, in medicine, an outcome of clinical procedure may be represented with a set of measurements [29]. In systems biology, any change in the environment or in the genome may be observed at a systemic level through a set of observed variables that describe the phenotype [5]. Hypotheses in chemical genomics need to relate the data on chemical structures with whole-genome observations [15]. In social sciences, data analysis may benefit from methods that can address a set of observed factors that describe human behavior [18].

To address such problems, we can benefit from the original definition of the subgroup discovery tasks, but take into consideration more complex data structures. In this article, we present an approach to subgroup discovery where the outcome is described with a set of *response variables*, or *output attributes*. For this purpose, we have developed an algorithm called MR-SD for *subgroup discovery from data with multi-dimensional responses*, or *multiple-response subgroup discovery* in short. MR-SD identifies a set of subgroups – collections of data instances from the training data set – where in each subgroup the data instances are similar in terms of the values of output attributes and are separable from the rest of the data in the input space.

[Figure 1 about here.]

Let us illustrate multiple-response subgroup mining through a hypothetical example. Figure 1 shows the sample values of two input (left column) and

two output attributes (right column). In Figure 1.a, the black circles mark a subset of data instances that are nicely clustered in both spaces. Clearly, these instances are similar in the output space, and comprise a well-defined neighborhood in the input space thus satisfying our constraint for an interesting subgroup. Similarly, the selected data instances in Figure 1.b are again clustered in the output space, and are linearly separable in the input space. This set could also constitute an interesting subgroup, but should use a different algorithm to report on separability in the input space. While the nearest neighbors algorithm may well-separate subgroup's instances from Figure 1.a, it would fail to do so for a subgroup from Figure 1.b where a linear classifier, like logistic regression, would succeed. In Figure 1.c, the selected instances are clustered well in the output space, but are not separable in the input space and therefore do not constitute an interesting subgroup.

The MR-SD algorithm described in the paper is based on the combination of unsupervised (clustering) and supervised (classification) techniques, and traverses a hierarchical clustering tree to obtain candidates for subgroups.

In the remainder of the paper, we first describe the related work. Then, we formally introduce the algorithm and a set of accompanying techniques for subgroup scoring and selection. We test the behavior of the proposed algorithm on two synthetic data sets and then describe an case study application in the area of analysis of data from European Social Survey. We conclude the paper with discussion and concluding remarks.

2 Related work

The type of the data analysis, where the inference algorithm aims at finding interesting data subsets which share unusual properties rather then producing the comprehensive model of the entire data set, was first proposed in 1996 by

Klösgen [21]. Referred to as *subgroup discovery*, one of the least complex of the subgroup's properties is related to a distribution of a single binary response variable. Indeed, initial efforts in subgroup discovery research and the majority of existing techniques deal with binary class-labeled data and for search of the subgroups propose various adaptations of rule learning algorithms. Prominent algorithms, that also vary in the utility of different rule scoring approaches [23] include SD [7], CN2-SD [24] and APRIORI-SD [20].

Rule-based approaches for subgroup discovery have also been adapted for discovery of more complex target concepts, including those that include multi-valued classes [1] (still using a single nominal class variable, but instead of two-valued encoding more than two classes) and a numerical response variable [6, 17]. Further, subgroup discovery approaches were proposed for the analysis of data with several binary response variables [39], and the analysis of inferred statistical models, the so-called exceptional model mining [25]. Instead of relating a set of attributes to a class, subgroup discovery can also relate to other properties in the data, like finding the subgroup-specific interactions between two variables [27, 32].

Neither of the approaches mentioned above can be directly applied to data sets with multi-dimensional response variables. In principle, one could binarize all the variables and use the cluster grouping approach [39]. In this way, however, additional parameters are required for this procedure, with potential loss of the ordering information for continuous attributes, yielding models that are harder to interpret. Alternatively, any of the existing algorithms for single-class subgroup discovery can be used for each response variable independently. Such approach would be computationally more expensive and would create a set of models instead of a single one thus hampering interpretability. Methods performing such analysis would also disregard any possible dependencies among

the response variables.

The inappropriateness of splitting the multi-dimensional response problem to several single-dimensional problems has already been exposed in early reports of algorithms that simultaneously predict a set of response variables. An excellent example of an algorithm that constructs a multi-target prediction model is the hierarchical clustering trees approach [14]. Clustering trees are a generalization of the decision trees that are able to treat several responses simultaneously. Instead of a single-class based attribute scoring (*e.g.*, entropy [30]), the split of the data at each node to a set of data subsets is scored according to within-subset instance similarity by comparing all corresponding pairs of data instances.

Clustering trees cannot be regarded as a subgroup discovery approach: the method aims to construct a global model of the data. Just like for the standard classification trees, the partition of the data strongly depends on the set of initial splits close to the root of the tree, and the algorithm may fail to uncover many of the interesting patterns due to the recursive nature of the tree construction algorithm. To remedy these shortcomings, Ženko proposed to use rule-based learning as a core algorithm for predictive clustering [34]. His method of predictive clustering rules (PCR) is an extension of the CN2 rule learning algorithm [3] that can model the distribution of a set of response variables instead of a single class. Just like CN2-SD [24], PCR uses a weighted covering rule-discovery technique, allowing the rules to refer to the similar sets of instances from the training set and thus encouraging the inference of overlapping rules. For rule scoring, PCR combines the estimate of accuracy of the rule for the nominal output variables with the decrease of variance of the continuous output variables.

Recently, a different rule-based approach for subgroup discovery was proposed by Hapfelmeier [13], addressing the data containing medical images described with a set of attributes. Authors first partitioned the data using k-

medoids clustering on images, and then inferred rules to assign cluster memberships using a standard subgroup discovery technique, the RSD algoritm [38], an extension of CN2-SD [24] for subgroup discovery in relational data sets. Partitioning clustering was also used by our own early approach to multi-response subgroup discovery [33], where data was clustered both in input and output space, and analysis of contingency tables was used to find potentially related clusters with a substantial number of overlapping instances. The major weaknesses of this approach are computational complexity (the method needed to traverse a set of combinations of parameters that determine the number of clusters in both spaces), implicit discovery of subgroups (intersection of clusters in input and output space), and reliance on external post-processing method for the construction of symbolic description for cluster membership.

3 Methods

An input to the proposed multi-response subgroup discovery algorithm (MR-SD) is a training data set E consisting of a random sample of n data instances e_i that are described with m_x input attributes and m_y output attributes. We will denote the two attribute sets with X (input attributes) and Y (output attributes). Each data instance e_i is therefore represented with a pair (x_i, y_i), where x_i is an attribute-value m_x-tuple and y_i an attribute-value m_y-tuple.

A *subgroup* G is a non-empty subset of the input data set E. The set of all possible subgroups $\mathcal{G}^\star = \mathcal{P}(E) \setminus \{\emptyset\}$ contains $2^n - 1$ subgroups, and is too large to be investigated exhaustively. MR-SD applies a heuristics where the data instances are first hierarchically clustered in the space of output attributes, and then candidate subgroups are obtained by traversal of the inferred clustering tree (see Figure 2). Each candidate subgroup is scored according to the separability of its data instances from all other training data instances using a selected

supervised data mining technique. Subgroups that exceed a specified score threshold are then ranked and reported to the data analyst. The proposed algorithm may infer subgroups that are similar in terms of covered instances, and we further propose a post-processing method to prune the subgroup set and identify best-scored subgroups with only a small mutual overlap. The details of the algorithm and its implementation are described below.

[Figure 2 about here.]

3.1 MR-SD Algorithm

MR-SD algorithm (Figure 3) starts with a hierarchical clustering of data instances in the space of output attributes Y. Instances in the same node of the clustering tree are therefore similar based on their output attribute values. The algorithm then verifies which of these clusters contain data instances that are separable in the input space X. Subgroup candidates are gathered by traversing the hierarchical clustering tree, testing if the instances in each of the node form a subgroup that can be reliably identified in the input space using a given supervised data mining technique. For each clustering tree node, it estimates the accuracy of a probabilistic classifier f_G when separating instances from the node (subgroup G) to their complement $G^C = E \setminus G$. We define the probabilistic classifier f_G as a mapping $f_G \colon E \rightarrow [0, 1]$ such that the value $f_G(e_i)$ is the estimation of probability $P(e_i \in G)$. This probability is estimated based only on the values of input attributes X. If the estimated accuracy of f_G exceeds a predefined threshold, we add the subgroup G to a set of interesting subgroups $\mathcal{G}$.

[Figure 3 about here.]

The algorithm is general in terms the clustering algorithm, dissimilarity measure, the supervised data mining method and its related scoring technique;

these are all parameters of the method. In our experiments, we used a standard agglomerative hierarchical clustering with Ward's linkage [36] and Manhattan distance on $[0, 1]$-scaled continuous attributes and discrete distance (0 for same-valued and 1 for different-valued) for nominal attributes.

We use cross-validation and estimate the area under receiver operating characteristics curve (AUC, [12]) to score separability of subgroup instances in the input attribute space. Given a probabilistic classifier f_G, AUC is equal to the probability that the classifier f_G distinguishes between a member of G and a member of G^C [12]. While any standard measure for classification accuracy can be used here, AUC's advantage is that its scale does not depend on the prior distribution of the class variable, in our case the frequencies of G and G^C. That is, with AUC, the threshold T_{Acc} will have the same meaning for subgroups of different sizes.

Subgroup discovery aims at identifying *descriptions* of interesting subgroups. The natural candidates for the supervised data mining technique are machine learning algorithms where the structure of the model can be easily communicated to the participating domain expert. In the case study included in this article, where the interpretation of results was essential, we used the naive Bayesian classifier and its nomogram-based model visualization [26].

3.2 Post-processing and subgroup selection

MR-SD's search heuristic identifies subgroups by traversing the hierarchical structure of clusters. Some of the discovered subgroups may therefore substantially overlap, that is, may share a large number of data instances. From the end-user's perspective it is desired that the algorithm would infer only a small subset of most representative subgroups. For this reason, and to avoid reporting similar subgroups albeit their high-scores, we define a subgroup post-

processing step. Its aim is to identify a subset of highest-scored subgroups with only a small overlap. In principle, if two subgroups substantially overlap, it is reasonable to report only the one with the higher score.

We measure the overlap of two subgroups G_i and G_j in terms of the proportion of jointly shared instances ($G_i \cap G_j$) among all instances covered by the two subgroups ($G_i \cup G_j$), and accordingly use the Jaccard coefficient of similarity [16]:

$$J(G_i, G_j) = \frac{|G_i \cap G_j|}{|G_i \cup G_j|} \tag{1}$$

We then create a network where each node represents a subgroup from $\mathcal{G}$. Two nodes in this network representing G_i and G_j are connected if $J(G_i, G_j) > T_J$. The threshold T_J is a user-defined parameter. Sufficiently large thresholds lead to fragmented networks composed of a number of connected network components. From each of the components, we select the highest-scored subgroup and include it in the final set of the subgroups $\mathcal{G}'$.

An example of subgroup similarity network is shown in Figure 4. The example demonstrates how the proposed post-processing step reduced a set of eight subgroups inferred by MR-SD to a subset of four, retaining the diverse subgroups with highest AUC scores.

[Figure 4 about here.]

4 Experiments on Synthetic Data Sets

We have studied the performance of the proposed multi-response subgroup discovery algorithm on artificially generated synthetic data sets. Synthetic data sets intentionally included a *target subgroup* and we tested if MR-SD can discover it under the presence of noise. We also used these data sets to compare MR-SD to predictive clustering rules [35, 34] and outline the differences between

these two approaches.

4.1 Data

We have generated two types of synthetic data sets, one with binary (D_B) and one with continuous (D_C) input attributes. The size of all data sets was set to $n = 200$. The data sets were generated so that to include a target subgroup of 40 (20%) data instances. The data sets included five output attributes ($Y_1, \ldots, Y_5$), for which the values were sampled either from a normal distribution $N(5, 1)$ for subgroup's instances or from $N(0, 1)$ for all other instances.

Data sets D_B included binary input attributes $X_1, X_2, \ldots, X_{m_x}$. Instances of the target subgroup G had first two attributes equal to 1, *e.g.* $X_1 = 1 \wedge X_2 = 1 \Leftrightarrow G$. Values of X_1 and X_2 for all other instances were chosen arbitrarily from discrete uniform distribution, but did not include the combination of $X_1 = 1 \wedge X_2 = 1$. Values of all other attributes $X_3, \ldots, X_{m_x}$ were sampled from discrete uniform distribution.

For D_C, the input attribute values were sampled from $N(0, 1)$. First, the subgroup G has been generated with sampling from the selected distribution but constraining the values of X_1 and X_2 to satisfy the relation $X_1^2 - X_2^2 - 1 \geq 0$. The same distribution was used in generating the complement, where X_1 and X_2 were constrained to $X_1^2 - X_2^2 - 1 < 0$.

The final data sets included either only two input attributes (data sets $D_{B,2}$, $D_{C,2}$) or ten input attributes (data sets $D_{B,10}$, $D_{C,10}$). For $D_{B,10}$ and $D_{C,10}$ eight input attributes are intentionally non-informative.

To distract otherwise clear separation of target subgroup in the output space we have added noise by choosing a proportion P_{noise} of data instances for which the values of their output attributes were randomly permuted across selected instances. Different noise levels from $P_{noise} = 0$ (no noise) to up to $P_{noise} = 0.3$

with step 0.01 were examined. The introduction of noise degraded both clustering properties of the data set in the output attribute space, and consequently the relations between input and output attributes (see Figure 5).

[Figure 5 about here.]

4.2 Evaluation Procedure

The synthetic data sets described above were used to test the MR-SD algorithm, and compare it to the PCR, the predictive clustering rules [34]. MR-SD was run with default parameters, using naive Bayesian classifier in case of discrete input attributes (D_B) and support vector machines with a polynomial kernel in case of continuously-valued input attributes (D_C). The threshold T_{Acc} was set to 0.75. PCR was run with the default parameters (multiplicative weighting scheme and search heuristics, disregarding other rules in the rule set). We tested the ability of the two procedures to uncover the target subgroup. The set of the subgroups $\mathcal{G}$ obtained by the algorithms was compared to the target subgroup, and the algorithm A was then scored according to the discovered subgroup $G_i \in \mathcal{G} = \mathcal{G}(A)$ which matched the target subgroup best in terms of the inclusion of the relevant data instances as measured by Jaccard coefficient of similarity:

$$score(A) = \max_{G_i \in \mathcal{G}} J(G_i, G). \tag{2}$$

The algorithms may fail to find any interesting subgroup, that is, none of examined subgroups is scored above T_{Acc}. The corresponding $score(A)$ is in such cases set to 0.

For each data set type and level of noise, the experiments were run 50 times. We report the averaged results and standard deviations.

4.3 Results and Discussion

The results of the experimental study on synthetic data sets are summarized in Figure 6. They show that the MR-SD behaves similarly to PCR if the number of input variables is small, the input attributes are discrete and if the function that maps the instances to a subgroup can be represented with a rule (Figure 6a). Adding irrelevant input attributes, at least to a degree studied here, does not hamper the performance of MR-SD, but may substantially degrade the performance of PCR (Figure 6b).

[Figure 6 about here.]

Results in Figures 6c and d confirm the applicability of MR-SD in case of continuous input attributes and more complex subgroup membership functions. The advantage of MR-SD is that it can use any suitable supervised data mining algorithm to characterize the subgroups in the input attribute space, and can hence accommodate for a wider variety of data types and subgroup description functions.

The scoring of subgroup discovery algorithms was based on the discovered subgroup that *best* matched the target subgroup G. For each data set, both algorithms proposed several subgroups. For data sets $D_{B,2}$ and $D_{B,10}$, MR-SD found more than ten subgroups, but which were similar to each other and have been represented with a single subgroup after the post-processing step. We have also observed that, regardless of the noise level, the subgroup most similar to the target subgroup was also the highest rated by the AUC-based scoring. While the evaluation scores in Figure 6 steadily decrease with increasing level of noise, the correct identification of the target subgroup despite noisy data speaks of the robustness of the proposed method. In contrast, PCR proposed two or three subgroups, of which the one matching the target was in most cases not the best-scored one.

For data sets $D_{C,2}$ and $D_{C,10}$, MR-SD discovers from 10 to 20 subgroups. The number of discovered subgroups increases with the increasing noise level. Regardless of noise, post-processing returned only one subgroup that matches the target subgroup for $D_{C,2}$, and returned two subgroups for $D_{C,10}$ (a target subgroup and a substantially different subgroup). The number of subgroups identified by PCR also increased with the level of noise, but failed to identify any subgroup that would substantially cover the target one. As the data sets $D_{C,2}$ and $D_{C,10}$ include the concept which cannot be represented in a rule-based language, adding noise actually helps PCR to propose subgroups that include few of the instances from the target subgroup.

Synthetic data sets studied in this section were crafted to test the capability of MR-SD to reveal the target subgroup, which was purposely designed to be "discoverable" by the selected supervised data mining technique. MR-SD had with this an advantage to PCR. Experiments show that MR-SD could be as successful as PCR at uncovering the subgroups that can be modeled with rules, but may outperform PCR when subgroup membership concepts cannot be modeled with classification rules.

The performance of MR-SD may depend on the choice of the supervised data mining algorithm, as a particular machine learning technique may either fail or succeed in modeling of data in the input space. To illustrate this point, consider the performance of MR-SD when varying this particular component of the overall algorithm (Figure 7). As proposed in this paper, the particular classification technique is a parameter of the method, and should be specified according to the user's knowledge of the domain.

[Figure 7 about here.]

5 Case Study: Analysis of the Data from European Social Survey

The European Social Survey (ESS) [18] is a biennial academically-driven social survey designed to describe attitudes, behavior and beliefs of European citizens. The aim of the case study was to apply MR-SD to a real-life data set and to observe if it can uncover interesting subgroups that would relate socio-demographic variables (input attributes) to various sets of output attributes, like attitudes and behaviors. Naive Bayesian classifier was used to characterize the subgroups in the input attribute space. Subgroups were requested to cover at least 5% but no more than 40% of instances. 5-fold cross validation was used to estimate AUC of the naive Bayesian classifier. The threshold T_{Acc} was set to 0.75, a lower bound for the acceptable AUC score [28]. In the post-processing stage, we set $T_J = \frac{1}{2}$.

5.1 Data

We have considered the data of the Slovenian survey from the year 2006. The survey included 1.476 persons and recorded over 300 different variables (see Table 1), including socio-demographic characteristics, the use of media, attributes recording social and public trust, and other.

From the data, we have excluded attributes that report on all the interviewer self-completion questions and test questions ($N = 22$). We have also excluded near-constant attributes ($N = 25$) having the same value for more than 90% data instances, and attributes having more than 10% of missing values ($N = 118$). We have then split the remaining ($N = 165$) attributes to the input and output set. The input set comprised 13 socio-demographic attributes including gender, age, marital status, and education level. The remaining attributes ($N =$

152) were divided into six non-overlapping blocks that correspond to six different sections from the questionary (Table 1). Our analysis was therefore comprised of six different tasks, each sharing the same set of input attributes, but using a different set of output attributes.

5.2 Results

Table 1 summarizes the results of subgroup discovery on the survey data set. MR-SD found interesting subgroups for five out of six modeling tasks. In all tasks where MR-SD initially proposed more than one subgroup, post-processing could substantially reduce the number of subgroups with minimal loss in the coverage. From the set of subgroups, we have selected two that we present in a detail below.

[Table 1 about here.]

5.2.1 Example Subgroup: Media and Social Trust

A subgroup of 457 (31%) data instances with AUC= 0.85. The naive Bayesian model for the subgroup membership is shown using the nomogram [26] in Figure 8. The nomogram depicts the influence of the five most informative input variables (age, level of education, ...). We can conclude that this is a subgroup of a younger, well-educated individuals. To summarize their characteristics with respect to the media and social trust, Table 2 ranks the output attributes according to the degree of difference between the distribution of their values in the subgroup and entire data set. For testing the differences in the distributions we used the t-test and reported the corresponding p–value which was further adjusted for the multiple comparisons using Bonferroni correction. According to this analysis, the individuals in the subgroup of younger, well-educated individuals very often use Internet and rarely listen to the radio.

[Figure 8 about here.]

[Table 2 about here.]

5.2.2 Example Subgroup: Attitudes and Timing of Life

A subgroup discovered for this task consists of 485 (33%) data instances and is well-characterized in the input attribute space (AUC=0.98). The characterization model in Figure 9 includes five most important input variables (legal marital status, age, ...). The most important properties of this subgroup in the output space are summarized in the Table 3. Analogously, the p–values have been computed using t–test for continuous and χ^2 test for categorical attributes and have been further adjusted with Bonferroni's correction.

According to the nomogram (Figure 9) this subgroup consists of younger, single individuals. The most subgroup-characteristic output variables (Table 3) turn out to measure similar properties ("not married", "not being parents"). Very high AUC score is therefore due to similarity of variables in input and output space. Although this subgroup does not represent any useful concept, the experiment nevertheless confirms the solid formal background of the proposed technique.

[Figure 9 about here.]

[Table 3 about here.]

6 Conclusion

In data rich domains, tools of data analytics often look for interesting data subgroups, rather than require a construction of a comprehensive model that would encompass entire data set. Methods of subgroup discovery have been tailored for this task. While most popular techniques in this field address data

with a single response variable, many of today's relevant problems from experimental research, industry, and socioeconomics may be described with the data whose description of the outcome is richer and, for example, includes a set of attributes instead of a single class. In the paper, we introduce a new technique that addresses such problems and is able to infer subgroups from data sets with multi-dimensional responses.

The task of multi-response subgroup discovery is to find subgroups of data sets similar in the description of the output and whose members can be characterized in the input attribute space. Our multi-dimensional subgroup discovery algorithm (MR-SD) directly addresses this task: candidate subgroups with data instances similar in outcomes are proposed by agglomerative clustering in the output space, and tested if they can be characterized in input space by means of the supervised data mining.

The major strength of the MR-SD are its reliance on standard, efficient and fast algorithms for clustering and machine learning, and on utility of a standard technique for subgroup scoring. Additional advantage is that any machine learning technique can be used for subgroup characterization. The particular choice would most likely depend on the type of the problem and input attributes, analyst's preference and familiarity with the technique and its ability to construct an interpretable model.

MR-SD has several potential weaknesses. Traversal of hierarchical clustering tree may yield a number of in composition very similar subgroup candidates, each of which is in the current implementation individually scored for its characterization in the input attribute space. The algorithm could be improved through sampling of the candidates, and could instead initially examine only those that are most different in composition. Identification of compositionally different subgroups is in the proposed version of the method performed after

the scoring, which is computationally more demanding, but – at least for the problem sets examined in this work – sufficiently efficient in terms of the runtime. Even in such settings, subgroup candidates from hierarchical clustering tree do not present the entire set of possible subgroups. Yet, hierarchical clustering provides means for an efficient heuristic search, aiming at minimizing the number of candidates to be scored in the input attribute space.

Being able to incorporate any classification algorithm for scoring of subgroups in the input space is an advantage of MR-SD, especially when compared to subgroup discovery methods that were developed around a particular modeling technique. This, however, adds another parameter to the method. In this paper, we did not explore means to automatic identification of suitable selection of supervised data mining technique, and see this as a potential future improvement of the approach.

Both experiments on synthetic data sets and on a real data set demonstrate that the method can discover relevant and interesting data subsets. Experiments on synthetic data sets also demonstrate the advantage of choosing an appropriate machine learning method for subgroup characterization. With increasing practical demand for multi-response subgroup mining techniques, further work should focus on practical applications and application-based optimizations of the approach.

References

[1] T. Abudawood and P. Flach. Evaluation measures for multi-class subgroup discovery. In *European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD '09)*, pp. 35–50, Berlin, Heidelberg, 2009. Springer-Verlag.

[2] M. Atzmueller, F. Puppe, and H.P. Buscher. Exploiting background knowledge for knowledge-intensive subgroup discovery. In *International Joint Conferences on Artificial Intelligence (IJCAI '05)*, pp. 647–652. Springer-Verlag, 2004.

[3] P. Clark and T. Niblett. The CN2 induction algorithm. *Machine Learning*, 3(4):261–283, 1989.

[4] M.J. del Jesus, P. González, F. Herrera, and M. Mesonero. Evolutionary fuzzy rule induction process for subgroup discovery: A case study in marketing. *IEEE Transactions on Fuzzy Systems*, 15(4):578–592, 2007.

[5] R. Franco-Duarte, L. Umek, B. Zupan, and D. Schuller. Computational approaches for the genetic and phenotypic characterization of a *saccharomyces cerevisiae* wine yeast collection. *Yeast*, 26(12):675–692, 2009.

[6] J.H. Friedman and N.I. Fisher. Bump hunting in high-dimensional data. *Statistics and Computing*, 9(2):123–143, 1999.

[7] D. Gamberger and N. Lavrač. Expert-guided subgroup discovery: Methodology and application. *Journal of Artificial Intelligence Research*, 17(1):501–527, 2002.

[8] D. Gamberger, N. Lavrač, A. Krstačić, and G. Krstačić. Clinical data analysis based on iterative subgroup discovery: Experiments in brain ischaemia data analysis. *Applied Intelligence*, 27(3):205–217, 2007.

[9] D. Gamberger, N. Lavrač, and G. Krstačić. Active subgroup mining: A case study in coronary heart disease risk group detection. *Artificial Intelligence in Medicine*, 28(1):27–57, 2003.

[10] D. Gamberger, N. Lavrač, F. Zelezny, and J. Tolar. Induction of comprehensible models for gene expression datasets by subgroup discovery methodology. *Journal of Biomedical Informatics*, 37(4):269–284, 2004.

[11] D.J. Hand, S. Padhraic, and H. Mannila. *Principles of data mining*. MIT Press, Cambridge, MA, USA, 2001.

[12] J.A. Hanley and B.J. Mcneil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, April 1982.

[13] A. Hapfelmeier, J. Schmidt, M. Mueller, S. Kramer, R. Perneczky, A. Kurz, and A. Drzezga. Interpreting PET scans by structured patient data: A data mining case study in dementia research. In *8th IEEE International Conference on Data Mining (ICDM '08)*, pp. 213–222, 2008.

[14] B. Hendrik, L. De Raedt, and J. Ramon. Top-down induction of clustering trees. In *15th International Conference on Machine Learning (ICML '98)*, pp. 55–63. Morgan Kaufmann, 1998.

[15] M.E. Hillenmeyer, E. Fung, J. Wildenhain, S.E. Pierce, S. Hoon, W. Lee, M. Proctor, R.P. St Onge, M. Tyers, D. Koller, R.B. Altman, R. W. Davis, C. Nislow, and G. Giaever. The chemical genomic portrait of yeast: Uncovering a phenotype for all genes. *Science*, 320(5874):362–365, 2008.

[16] P. Jaccard. Nouvelles recherches sur la distribution florale. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, 2(44):223–270, 1908.

[17] A.M. Jorge, F. Pereira, and P.J. Azevedo. Visual interactive subgroup discovery with numerical properties of interest. In *9th International Conference on Discovery Science (DS '06)*, Lecture Notes on Computer Science, pp. 301–305. Barcelona,Spain, Springer-Verlag, 2006.

[18] R. Jowell and the Central Co-ordinating Team. European social survey 2006/2007. Technical Report, London: Centre for Comparative Social Surveys, City University, 2007.

[19] B. Kavšek and N. Lavrač. Rule induction for subgroup discovery: A case study in mining UK traffic accident data. In *5th International Multi-Conference on Information Society (IS '02)*, pp. 127–130. Springer-Verlag, 2002.

[20] B. Kavšek and N. Lavrač. APRIORI-SD: Adapting association rule learning to subgroup discovery. *Applied Artificial Intelligence*, 20(7):543–583, 2006.

[21] W. Klösgen. *EXPLORA: A multipattern and multistrategy discovery assistant*, pp. 249–271. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996.

[22] N. Lavrač, P. Flach, B. Kavšek, and L. Todorovski. Adapting classification rule induction to subgroup discovery. In *2nd IEEE International Conference on Data Mining (ICDM '02)*, pp. 266–273, 2002.

[23] N. Lavrač, P. Flach, and B. Zupan. Rule evaluation measures: A unifying view. In *9th International Workshop on Inductive Logic Programming (ILP '99)*, volume 1634 of *Lecture Notes on Computer Science*, pp. 174–185. Springer, 1999.

[24] N. Lavrač, B. Kavšek, P. Flach, and L. Todorovski. Subgroup discovery with CN2-SD. *Journal of Machine Learning Research*, 5:153–188, 2004.

[25] D. Leman, A. Feelders, and A. Knobbe. Exceptional model mining. In *European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD '08)*, volume 5212 of *Lecture Notes in Computer Science*, pp. 1–16. Springer, 2008.

[26] M. Možina, J. Demšar, M.W. Kattan, and B. Zupan. Nomograms for visualization of naive Bayesian classifier. In *11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04)*, pp. 337–348, 2004.

[27] M. Mueller, R. Rosales, S. Steck, S. Krishnan, B. Rao, and S. Kramer. Subgroup discovery for test selection: A novel approach and its application to breast cancer diagnosis. In *8th International Symposium on Intelligent Data Analysis (IDA '09)*, pp. 119–130, 2009.

[28] R. Murphy-Filkins, D. Teres, S. Lemeshow, and D.W. Hosmer. Effect of changing patient mix on the performance of an intensive care unit severity-of-illness model: How to distinguish a general from a specialty intensive care unit. *Critical Care Medicine*, 24(12):1968–73, 12 1996.

[29] NAMCS. National Center for Health Statistics. Surveys and Data Collection Systems. `http://www.cdc.gov/nchs/surveys.htm`, June 2010.

[30] J.R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[31] C. Romero, P. González, S. Ventura, M.J. del Jesus, and F. Herrera. Evolutionary algorithms for subgroup discovery in e-learning: A practical application using moodle data. *Expert Systems with Applications*, 36(2):1632–1644, 2009.

[32] X. Su, C.L. Tsai, H. Wang, D.M. Nickerson, and B. Li. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10:141–158, 2009.

[33] L. Umek, B. Zupan, M. Toplak, A. Morin, J.H. Chauchat, G. Makovec, and D. Smrke. Subgroup discovery in data sets with multi-dimensional

responses: A method and a case study in traumatology. In *12th Conference on Artificial Intelligence in Medicine (AIME '09)*, pp. 265–274, 2009.

[34] B. Ženko. Learning predictive clustering rules, Ph.D. dissertation, 2007.

[35] B. Ženko and J. Struyf. Learning predictive clustering rules. In *4th International Workshop on Knowledge Discovery in Inductive Databases: Revised Selected and Invited Papers (KDID '05)*, volume 3933 of *Lecture Notes on Computer Science*, pp. 234–250. Springer, 2005.

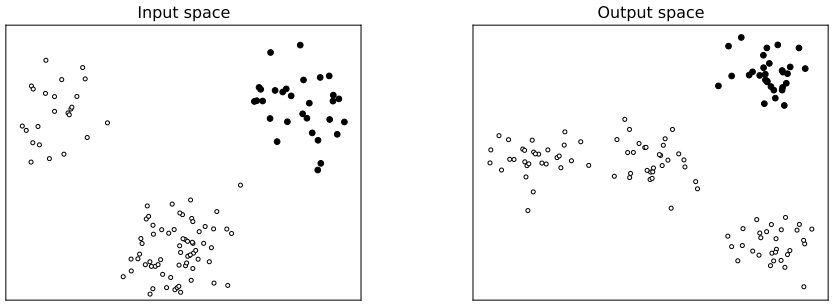
[36] J.H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[37] S. Wröbel. An algorithm for multi-relational discovery of subgroups. In *1st European Symposium on Principles of Knowledge Discovery in Databases (PKDD' 97)*, pp. 78–87, 1997.

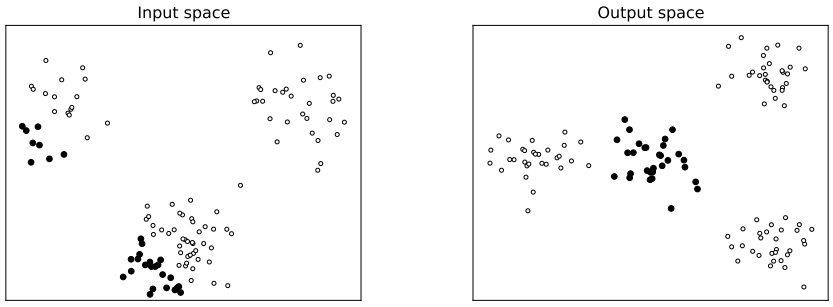
[38] F. Zelezny and N. Lavrač. Propositionalization-based relational subgroup discovery with RSD. *Machine Learning*, 62(1-2):33–63, 2006.

[39] A. Zimmermann and L. De Raedt. Inductive querying for discovering subgroups and clusters. In *European Workshop on Inductive Databases and Constraint Based Mining*, volume 3848 of *Lecture Notes on Computer Science*, pp. 380–399. Springer-Verlag, 2005.

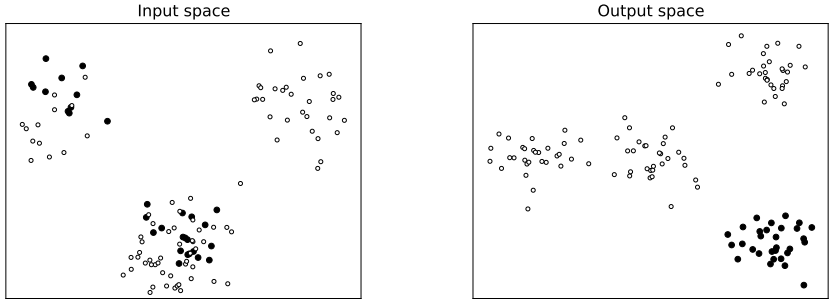
List of Figures



(a) Selected data instances (black circles) perfectly cluster in input and output space.



(b) Selected data instances (black circles) perfectly cluster in the output space, and can be in the input space separated from the rest of the data set using some linear-based classification method.



(c) The selected data instances (black circles) form a cluster in the output space, but cannot be separated from the rest of the data set in the input space.

Figure 1: Examples showing sets of data instances that are similar in terms of output attributes (data projections on the right) and are thus candidates for the subgroups. They, however, show different degrees and types of separability from the rest of the data instances in the input space (data projections on the left).

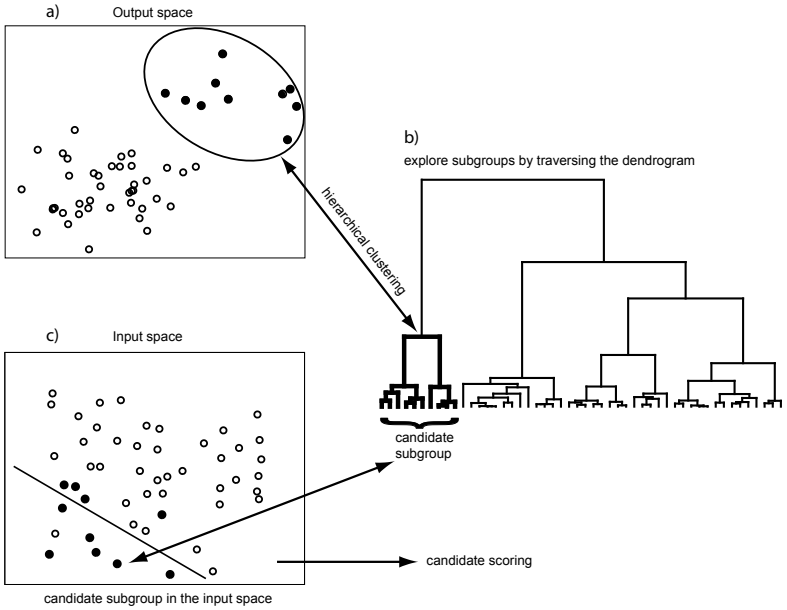


Figure 2: An outline of the multi-response subgroup discovery algorithm. The algorithm clusters the data in the space of output attributes (b) and traverses a resulting clustering tree to identify the candidates for the subgroups. An output-space projection of instances in candidate subgroups is shown in (a). Subgroups are scored according to the separability of the subgroup's instances from all other instances in the training set in the input space (c).

- Input:
  - A training data set $E = \{e_1, \ldots, e_n\}$
  - A set of input X and output Y attributes, such that $(X, Y)(e_i) = (x_i, y_i)$
  - A hierarchical clustering method for identification of groups of similar instances in the output space (dissimilarity measure d_Y using values of output attributes Y, linkage)
  - A supervised data mining technique for classification of data instances in the space of input attributes
  - A method for accuracy scoring of supervised data mining
  - A set of user-defined constraints: minimal and maximal subgroup size (m and M, respectively), accuracy score threshold T_{Acc}
- Output:
  - A set of subgroups $\mathcal{G} = \{G_1, \ldots, G_k\}$, where each subgroup G_i is composed of instances that represent a cluster in Y-space, are separable in X-space, and satisfy a set of user-defined constraints
- Inference procedure:
  1. Perform the hierarchical clustering of the training data using d_Y and the selected linkage method
  2. $\mathcal{G} \leftarrow \emptyset$
  3. **For** each node N of the clustering tree with at least m and at most M instances **do**:
     (a) $G \leftarrow$ set of data instances in the node N
     (b) Score the accuracy $Acc(G)$ of the supervised data mining technique for the task of separating instances in G from all other instances in the training set E using the values of input attributes X
     (c) **If** $Acc(G) \geq T_{Acc}$ **then** $\mathcal{G} \leftarrow \mathcal{G} \cup \{G\}$

Figure 3: MR-SD Algorithm.

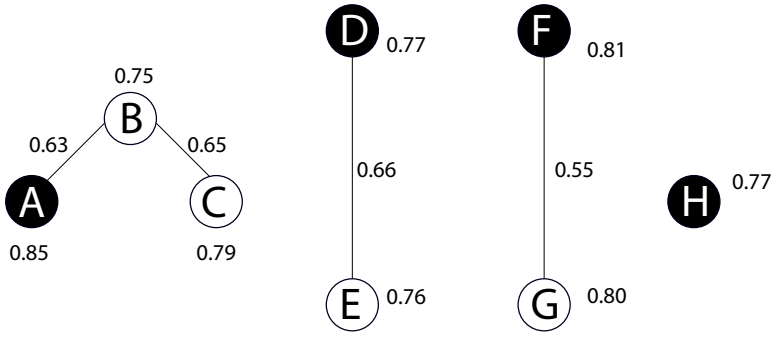
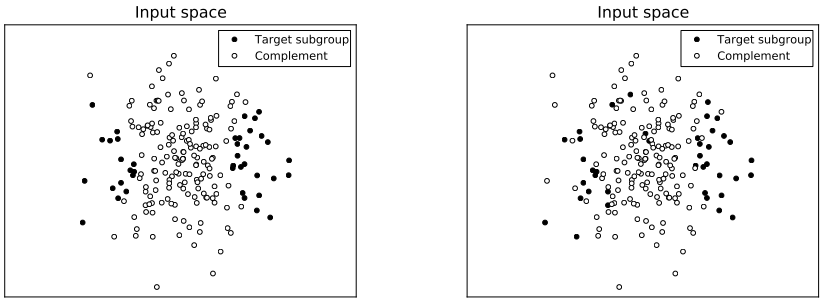


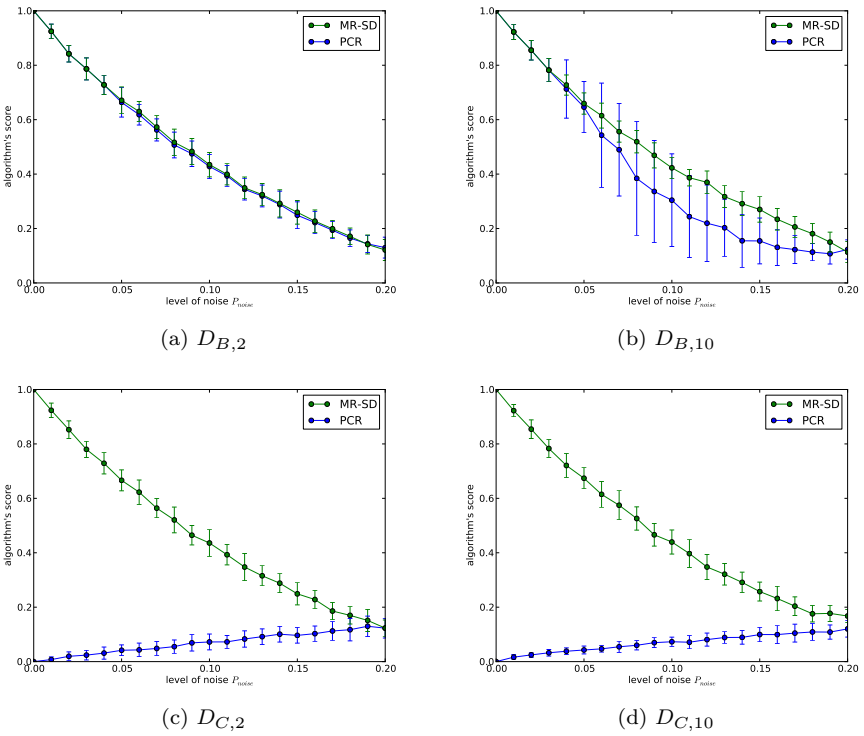
Figure 4: Subgroup similarity network for media and social trust data from European Social Survey (see Table 1). Nodes in the network represent eight discovered subgroups (A, B, ..., H); the edges connect two sugroups if their similarity measured by Jaccard coefficient exceeds $T_J = \frac{1}{2}$. Values adjacent to nodes are subgroups' AUC scores. Edge labels report on corresponding similarity score. This network has four connected components (ABC, DE, FG, H) from which a post-processing step selects a single subgroup with the highest AUC score. The original set of eight subgroups is after this step reduced to a subset of four subgroups (black colored nodes A, D, F, and H).



(a) $D_{C,2}$ without noise

(b) $D_{C,2}$ with maximal examined noise

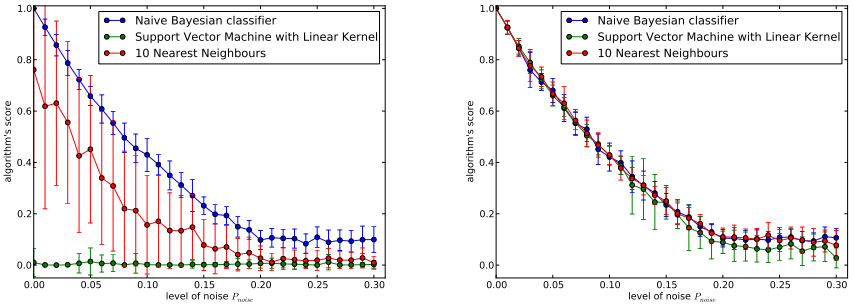
Figure 5: Instances from synthetic data set $D_{C,2}$ in the input space before (a) and after adding noise to the 30% of instances in the output space (b). After performing hierarchical clustering in the output space the task of the MR-SD algorithm is reduced to a separation between instances within and outside the candidate subgroup (e.g., discrimination between instances indicated with black and white circles). Notice that only adding noise does not change the data in the input space, but rather affects the identification and composition of the candidate subgroups.



(a) $D_{B,2}$ (b) $D_{B,10}$

(c) $D_{C,2}$ (d) $D_{C,10}$

Figure 6: Experimental results on synthetic data sets. Algorithms were scored according to the degree of uncovering the target subgroup (*score*).



(a) $D_{C,10}$ (b) $D_{B,10}$

Figure 7: The impact of selection of supervised data mining technique to the performance of MR-SD algorithm. Three different classification methods were used in subgroup discovery from the two synthetic data sets ($D_{C,10}$, $D_{B,10}$). In some situations, a wrong selection of a classification algorithm can lead to poor performance, such as choice of linear SVM in (a). In the same data set nearest neighbor classifier yielded less stable performance than naive Bayesian classifier. In contrast, the choice of the classification algorithms for $D_{B,10}$ data set (b) has no significant impact to the performance of MR-SD.

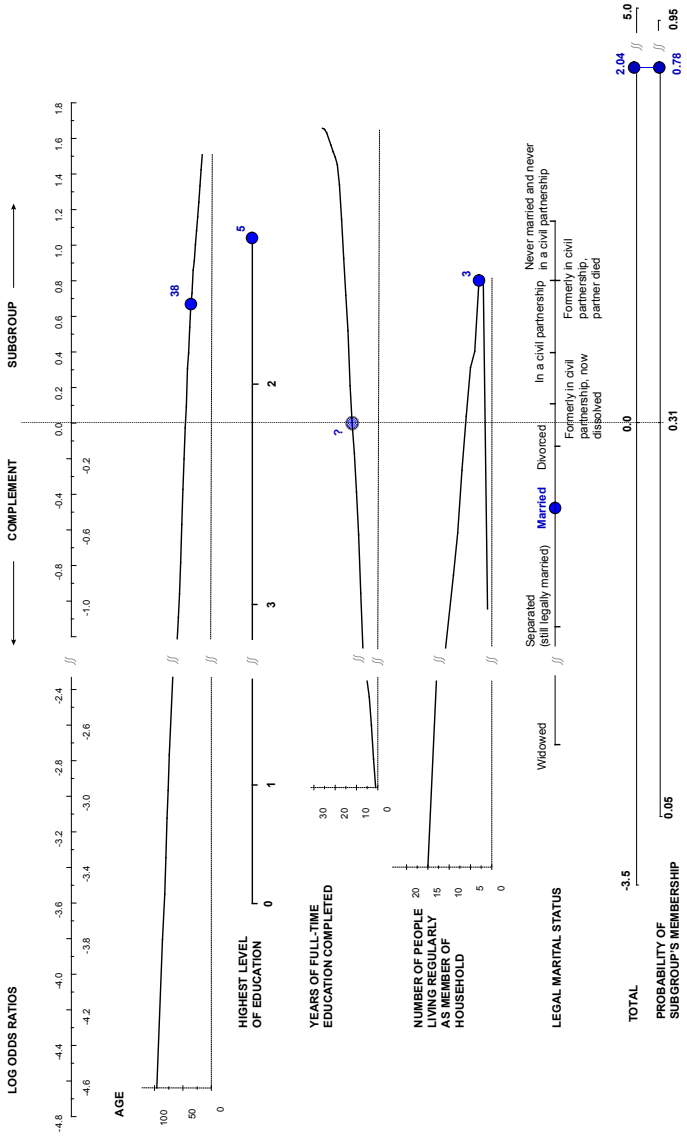


Figure 8: The naive Bayesian nomogram for the computation of the probability that a data instance belongs to the discovered subgroup in media and social trust task. The dots in the nomogram show values of a particular data instance: a 38-years old married individual with highest level of education who lives in a household together with two other people. The nomogram indicates that this individual is classified to the subgroup with 78% probability.

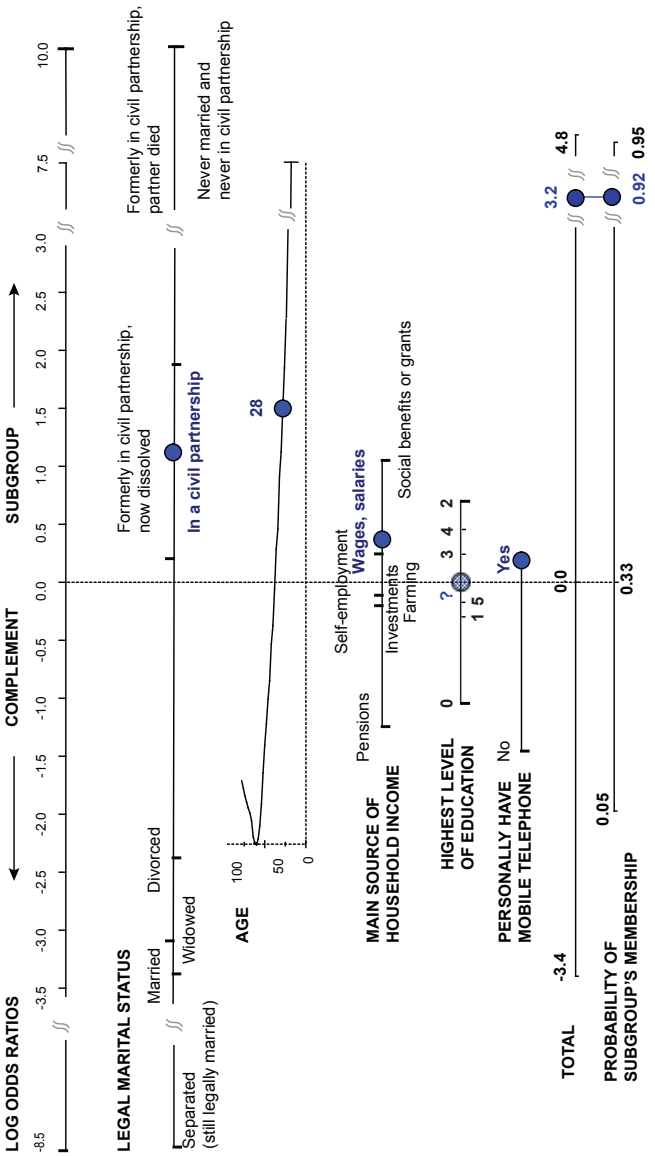


Figure 9: The naive Bayesian nomogram for the computation of the probability that a data instance belongs to the discovered subgroup in attitudes and timing of life task. Dots on the nomogram show values of a particular data instance: a 28-years old living in a civil partnership, who gets paid from wages and owning a mobile phone. The nomogram indicates that this individual is classified to a subgroup with a high 92% probability.

List of Tables

block of output attributes	$\|Y\|$	examples	$\|\mathcal{G}\|$ cov	$\|\mathcal{G}'\|$, cov	max AUC	average AUC
media, social trust	8	usage of internet	8	4	0.85	0.80
		reading newspaper	0.70	0.70		
		trusted in people				
politics	31	trust in country's parliament	1	1	0.75	0.75
		voted on last elections	0.29	0.29		
		signed petition				
subjective well-being,	20	pray (how often)	3	1	0.80	0.80
religion		take part in social activities	0.12	0.07		
		level of happiness				
attitudes,	34	age to become adult	13	6	0.98	0.80
timing of life		willful not to have children	0.58	0.48		
		being grandparent being old				
personal and social	38	enjoy life	6	1	0.78	0.77
well-being		feel anxious	0.34	0.16		
		have a lot of energy				
human values	21	important to seek fun	0	0	N/A	N/A
		important to try different things	0.00	0.00		
		important to behave properly				

Table 1: Overview of the results from six analyzed data sets. The reported are the number of output attributes ($|Y|$), list of several concrete examples of output attributes (abbreviated titles of survey questions), number of discovered subgroups ($|\mathcal{G}|$), number of subgroups remaining after post-processing ($|\mathcal{G}'|$) with the corresponding relative coverage of examples (cov), and maximal and average AUC score for subgroups for subgroups in $\mathcal{G}'$.

output attribute	adjusted p–value	mean $\pm$ std on E	mean $\pm$ std on G
personal usage of Internet e-mail, www (days per week)	$2.4 \cdot 10^{-85}$	3.5 ± 2.9	6.3 ± 1.2
radio listening on average weekday	$1.0 \cdot 10^{-34}$	3.5 ± 2.6	1.9 ± 1.6
most people can be trusted (from 0 to 7)	$7.4 \cdot 10^{-10}$	4.0 ± 2.7	4.9 ± 2.3
TV watching (news, politics) on average weekday	$2.2 \cdot 10^{-9}$	1.7 ± 1.2	1.3 ± 1.0
most people are fair (from 0 to 7)	$2.0 \cdot 10^{-3}$	4.9 ± 2.5	5.3 ± 2.1
TV watching on average weekday	$6.0 \cdot 10^{-3}$	3.4 ± 2.4	3.1 ± 2.1
most people are helpful (from 0 to 7)	$9 \cdot 10^{-3}$	4.5 ± 2.4	5.0 ± 2.1

Table 2: The most characteristic output attributes for subgroup's description in media and social trust task. Attributes with the adjusted p–value less than 0.01 are shown in the table. All output attributes in this task are continuous, their distribution is described in terms of their means and standard deviations on the subgroup G and on the entire data set E.

output attribute Y	adjusted p–value	distribution on E	distribution on G
have you been married	$2.6 \cdot 10^{-119}$	$P(Y = yes) = 0.65$	$P(Y = yes) = 0.03$
have you ever given birth or fathered a child	$1.7 \cdot 10^{-101}$	$P(Y = yes) = 0.70$	$P(Y = yes) = 0.13$
have you ever lived with your partner for more than 3 months	$3.3 \cdot 10^{-78}$	$P(Y = yes) = 0.70$	$P(Y = yes) = 0.29$
have you paid employment more than 20 hours per week	$3.8 \cdot 10^{-15}$	$P(Y = yes) = 0.75$	$P(Y = yes) = 0.55$
approve if person lives with partner not married to (from 0 to 7)	$6.5 \cdot 10^{-7}$	$mean = 3.45$ $sd = 0.93$	$mean = 3.70$ $sd = 0.81$

Table 3: The most characteristic output attributes for subgroup's description in attitudes and timing of life task. The table summarizes the five most important according the adjusted p–value. For this task, the most characteristic attributes are either binary or continuous. Their distribution is represented either with $P(Y = yes)$ or with the mean and standard deviation.